SNP Function Portal: A Web Database for Exploring the Functional Implications of SNP Alleles

Pinglang Wang[1], Manhong Dai[1], Weijian Xuan[1], Richard C McEachin[2], Anne U Jackson[3], Laura J Scott[3], Brian Athey[2], Stanley J Watson[1], Fan Meng[1,2]

[1]Molecular and Behavioral Neuroscience Institute and Department of Psychiatry, U of M, Ann Arbor, MI 48109

[2]National Center for Integrative Biomedical Informatics, U of M, Ann Arbor, MI 48109

[3]Biostatistics Department, School of Public Health, U of M, Ann Arbor, MI 48109

Abstract

Finding the potential functional significance of SNPs is a major bottleneck in understanding genome-wide SNP scanning results, as the related functional data are distributed across many different databases. The SNP Function Portal is designed to be a clearing house for all public domain SNP functional annotation data, as well as in-house functional annotations derived from different data sources. It currently contains SNP functional annotations in six major categories including genomic elements, transcription regulation, protein function, pathway, disease and population genetics. Besides extensive SNP functional annotations, the SNP Function Portal includes a powerful search engine that accepts different types of genetic markers as input and identifies all genetically related SNPs based on the HapMap Phase II data as well as the relationship of different markers to known genes. As a result, our system allows users to identify the potential biological impact of genetic markers and complex relationships among genetic markers and genes, and it greatly facilitates knowledge discovery in genome-wide SNP scanning experiments.
http://brainarray.mbni.med.umich.edu/Brainarray/Database/SearchSNP/snpfunc.aspx

Keywords: SNP, Annotations, Linkage Disequilibrium

Introduction

A Single Nucleotide Polymorphism (SNP) is a DNA sequence variation at a single nucleotide level. It is estimated that SNPs occur once per 100~300 bases in the human genome. The dramatic increase in genotyping efficiency in the last couple of years has made large-scale high density genome-wide SNP association analysis practical for many research groups. Genotype data have much more complex and indirect relationships with genes and proteins. Most SNPs are not even in the coding sequences of genes. They may influence biological processes in many conceivable ways: reduce transcription factor binding affinity to the promoter region, alter a microRNA binding site, change mRNA stability, modify the RNA splicing pattern, destroy an internal ribosomal binding site, etc. Given the complexity of the way that a SNP allele may influence the function of a protein, it is highly desirable to have a comprehensive database where researchers can easily access the most up-to-date SNP functional annotations. In addition, neighboring SNPs usually show different degree of linkage disequilibrium (LD).

Consequently, although a SNP allele itself does not cause any functional difference, a tightly linked nearby allele may be the causative allele. Although there are several efforts on the functional annotation of SNPs, the coverage of existing commercial or public domain efforts is far from complete. The main goal of this work is to build a comprehensive SNP function portal to facilitate the understanding of functional implications of SNP alleles identified in genome-wide association studies. We integrate annotation from different databases and generate functional annotations based on various existing sequence and structure analysis algorithms. We also provide annotation of SNPs related to high level biological functions. We integrate a powerful SNP search function that utilizes the LD data from the HapMap project in the SNP function search process. The portal accepts generic markers including SNPs, genes, microsatellite markers and cytogenetic bands as input.

To meet the requirements of different users, we currently provide a web service for identifying all genetically related SNPs, as well as batch annotation data download in multiple formats (text, Excel spreadsheet, etc.).
The SNP Function Portal greatly increases researcher's efficiency at SNP function exploration and it will be continuously improved by adding more features and functional annotations

System & Method

The SNP Function Portal currently has three main modules: 1) a powerful SNP search function that maps input genetic markers to all physically or genetically related SNPs satisfying user's criteria, based on the HapMap II and dbSNP data 2) a SNP function data integration pipeline that obtains updated annotation data from external sources and generates annotations using existing sequence and structure analysis programs, and 3) a web interface that receives user input, generates a summary report and provides flexible browsing, filtering, sorting and downloading capabilities. We collect SNP functional annotations from various sources and organize them into six major categories: Genome, Transcript, Protein, Pathway/Ontology, Disease (Table 1). These categories form a core framework to encapsulate existing and new annotations in our database. The overview of data sources, flow and annotation organization is shown in Figure 1.

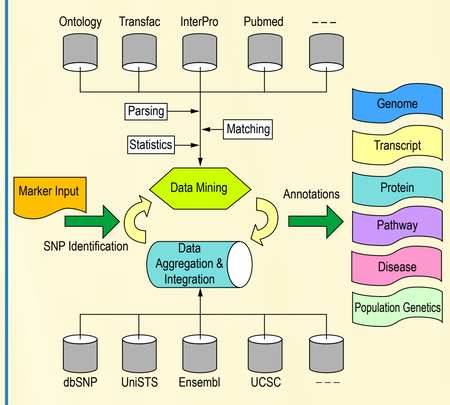

Figure 1. Overview of Data Sources, Organization and Flows

Genome Level	Conserved Genomic Region
	Repetitive Sequence
Transcript Level	CpG Island
	DNase 1 Hypersensitive Site
	Histone Acetylation Site
	Known Transcription Factor Binding Site
	Predicted Transcription Factor Binding Site
	RNA Polymerase Binding Region
	Splicing donar/acceptor sites
	Intronic and Exonic Splice Enhancer/Silencer
	Branch Site Recognition Sequence
	MicroRNA Transcript
	MicroRNA Binding Site
	RNA Stability
Protein Level	Internal Ribosomal Entry Site
	Known or Potential Modification Site
	Key Residues Influence Protein Domain Structure
	Protein Location Motif: Signal Peptide, Nuclear Localization Signal
	Conserved Protein Domain
	Activity/Binding Center Of Protein
	Protein-Protein Interaction Interface
Pathway/ Ontology Level	Gene Ontology
	KEGG Pathway
	BioCarta Pathway
	Molecular Interaction Network
Disease Level	OMIM annotation
	Literature description
	Association Statistics
	Linkage Statistics
	SNP-SNP Interaction Statistics
	Expression Profile
Population Genetics Level	Linkage Disequilibrium Scores
	Haplotype
	Ancestral Allele
	Tajima's D
	Fst Value
	Heterozygosity
	Allele Frequency
	Map Weight

Table 1. Annotation Categories

SNP Search Pipeline

Our search engine will automatically search and identify all the SNPs located in those genomic regions. One may also select gene neighbor to include SNPs in the context of the genes and their promoter regions for the SNPs in their input list. Our search engine will first search for all the Entrez genes that the input SNPs belong to, and then include all the SNPs in the genes as well as their 5' upstream regions users select. Our search engine will automatically search and identify all the SNPs located in those genomic regions. One may also select gene neighbor to include SNPs in the context of the genes and their promoter regions for the SNPs in their input list. Our search engine will first search for all the Entrez genes that the input SNPs belong to, and then include all the SNPs in the genes as well as their 5' upstream regions users select, as shown in Figure 2

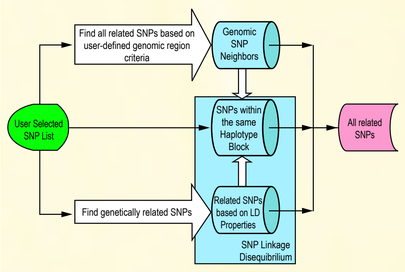

Figure 2. Functional Properties

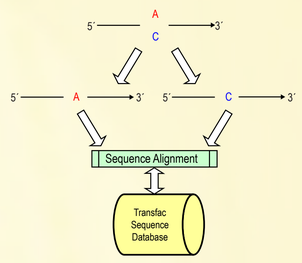

Figure 3. Analysis of SNP alleles against transcription factor binding sites

Custom Annotations

In addition to integrating public SNP annotations, we generate custom annotations through sequence analysis of SNP variations. 1) We analyze the sequences containing SNP alleles against Transfac database and try to identify the transcription factor binding sites they may affect (Figure 3). 2) We analyze the impacts of nonsynonymous SNPs on protein domain structure through protein sequence alignment with InterproScan 3) We also map SNPs to the Entrez genes and subsequently to different functional categories such as GO, KEGG, BioCarta, etc. 4) Basing on SNPs' genomic locations, we map them to diseases in OMIM through genetic marker linkage. 5) We also provide convenient external web links through reference SNP IDs for the databases not integrated in our SNP portal.

Web Portal

The initial web query interface taking various user criteria is demontrated in Figure 4. Samples of annotations for output SNPs are shown in Figure 5 and 6. In addition, we provide web service for researchers who only want to retrieve the list of related SNPs through our search engine described in Figure 2.

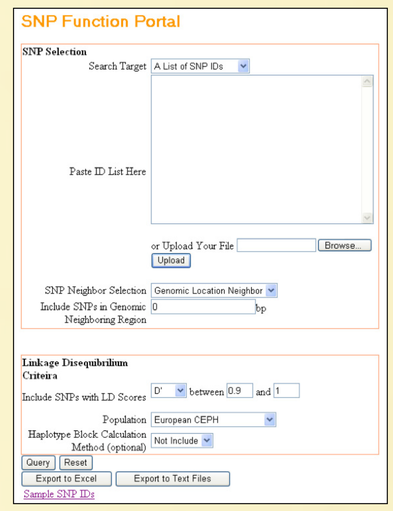

Figure 4. SNP Portal Search Interface

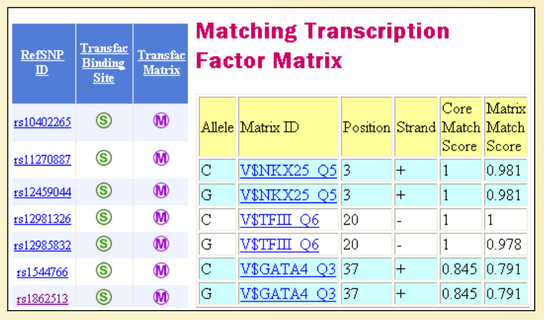

Figure 5. Transcription Factor Binding Site Annotation

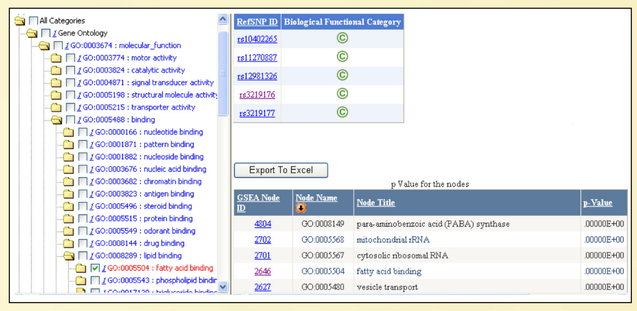

Figure 6. Gene Ontology Annotations

Summary

In summary, the SNP Function Portal is a one-stop solution for exploring the potential functional implications of different types of genetic markers through a powerful HapMap II-based search function. Although we certainly need to add more SNP functional annotation data sets, it is already the most powerful SNP annotation web function in the public domain. In the next phase of development, we plan to consider the use of haplotypes in protein domain analysis.
The current version is based on the effect of individual SNPs, as the current HapMap data available at that time did not include enough coding region nonsynonymous SNPs for generating meaningful haplotype-based protein domain analysis data for most of the proteins.

Acknowledgement

P. Wang, M. Dai, W. Xuan, S. J. Watson and F. Meng are members of the Pritzker Neuropsychiatric Disorders Research Consortium, which is supported by the Pritzker Neuropsychiatric Disorders Research Fund L.L.C.. This work was partly supported by the National Center for Integrative Biomedical Informatics under National Institutes of Health Grant #U54 DA021519-01A1.

This paper has been accepted for podium presentation at ISMB 2006 and will appear in the Bioinformatics" journal.